DOCUMENT RESUME

ED 390 918 TM 024 322

AUTHOR	Shepard, Lorrie A.; And Others
TITLE	Effects of Introducing Classroom Performance Assessments on Student Learning.
INSTITUTION	Colorado Univ., Boulder.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CSE-TR-394
PUB DATE	Feb 95
CONTRACT	R117G10027
NOTE	35p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Achievement Tests; *Classroom Techniques; Comparative Analysis; Elementary School Students; *Elementary School Teachers; Grade 3; Learning; *Performance Based Assessment; Pretests Posttests; Primary Education; Research Utilization; *Teacher Made Tests; *Test Construction; Test Use
IDENTIFIERS	Maryland School Performance Assessment Program

ABSTRACT

The effects of performance assessments on student learning were examined in a year-long project to help teachers in 13 third-grade classrooms begin to use performance assessments as part of regular instruction in reading and mathematics. The essential research question was whether students learned more or developed qualitatively different understandings because performance assessments were introduced. Achievement results for the approximately 335 students were compared to the performance of third-grade students in the same schools the year before and to third-grade performance in matched control schools. Researchers worked with the teachers throughout the year to help them develop performance assessments congruent with their own instructional goals. Standardized achievement tests and some items from the Maryland State Department of Education's performance assessment were used to measure achievement. Overall, the predominant finding was one of no difference or no gains in student learning following the year-long effort to introduce performance assessment. Researchers indicated that they saw qualitative changes in performance, but it must be acknowledged that any demonstrated benefits in achievement were small and ephemeral. (Contains 10 tables, 4 figures, and 18 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

ED 390 918

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Effects of Introducing Classroom Performance Assessments on Student Learning

CSE Technical Report 394

Lorrie A. Shepard, Roberta J. Flexer, Elfrieda H. Hiebert, Scott F. Marion, Vicky Mayfield, and Timothy J. Weston

CRESST/University of Colorado at Boulder

► **UCLA Center for the Study of Evaluation**

In collaboration with:

- ► University of Colorado
- ► NORC, University of Chicago
- ► LRDC, University of Pittsburgh
- ► University of California, Santa Barbara
- ► University of Southern California
- ► The RAND Corporation

TM024322

BEST COPY AVAILABLE

Effects of Introducing Classroom Performance Assessments on Student Learning

CSE Technical Report 394

Lorrie A. Shepard, Roberta J. Flexer, Elfrieda H. Hiebert,
Scott F. Marion, Vicky Mayfield, and Timothy J. Weston
CRESST/University of Colorado at Boulder

February 1995

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1521
(310) 206-1532

Copyright © 1995 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

PREFACE

The current intense interest in alternative forms of assessment is based on a number of assumptions that are as yet untested. In particular, the claim that authentic assessments will improve instruction and student learning is supported only by negative evidence from research on the effects of traditional multiple-choice tests. Because it has been shown that student learning is reduced by teaching to tests of low-level skills, it is theorized that teaching to more curricularly defensible tests will improve student learning (Frederiksen & Collins, 1989; Resnick & Resnick, 1992). In our current research for the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) we are examining the actual effects of introducing new forms of assessment at the classroom level.

Derived from theoretical arguments about the anticipated effects of authentic assessments and from the framework of past empirical studies that examined the effects of standardized tests (Shepard, 1991), our study examines a number of interrelated research questions:

1. What logistical constraints must be respected in developing alternative assessments for classroom purposes? What are the features of assessments that can feasibly be integrated with instruction?

2. What changes occur in teachers' knowledge and beliefs about assessment as a result of the project? What changes occur in classroom assessment practices? Are these changes different in writing, reading, and mathematics, or by type of school?

3. What changes occur in teachers' knowledge and beliefs about instruction as a result of the project? What changes occur in instructional practices? Are these changes different in writing, reading, and mathematics, or by type of school?

4. What is the effect of new assessments on student learning? What picture of student learning is suggested by improvements as measured by the new assessments? Are gains in student achievement corroborated by external measures?

5. What is the impact of new assessments on parents' understandings of the curriculum and their children's progress? Are new forms of assessment credible to parents and other "accountability audiences" such as school boards and accountability committees?

This report is of one of a set of papers that were presented at the 1994 annual meeting of the American Educational Research Association and summarize current project findings.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*, 232-238.

EFFECTS OF INTRODUCING CLASSROOM PERFORMANCE ASSESSMENTS ON STUDENT LEARNING[1,2]

Lorrie A. Shepard, Roberta J. Flexer, Elfrieda H. Hiebert, Scott F. Marion, Vicky Mayfield, and Timothy J. Weston

CRESST/University of Colorado at Boulder

Arguments favoring the use of performance assessments make two related but distinct claims. Performance assessments are expected first to provide better measurement and, second, to improve teaching and learning. Although any measuring device is corruptible, performance measures have the potential for increased validity because the performance tasks are themselves demonstrations of important learning goals rather than indirect indicators of achievement (Resnick & Resnick, 1992). According to Frederiksen and Collins (1989), Wiggins (1989), and others, performance assessments should enhance the validity of measurement by representing the full range of desired learning outcomes; by preserving the complexity (and ambiguity) of disciplinary knowledge domains and skills; by representing the contexts in which knowledge must ultimately be applied; and by adapting the modes of assessment to enable students to show what they know. The more assessments embody authentic criterion performances, the less we have to worry about drawing inferences from test results to remote constructs.

The expected positive effects of performance assessments on teaching and learning follow from their substantive validity. If assessments capture learning expectations fully, then when teachers provide coaching and practice to improve scores, they will directly improve student learning without corrupting the meaning of the indicator. Resnick and Resnick (1992), Frederiksen and Collins (1989), and

---

[1] Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 1994.

[2] We thank the Maryland Department of Education for allowing us to use tasks from the Maryland School Performance Assessment Program as outcome measures for the study. We also thank the Riverside Publishing Company for permission to use portions of the 2nd-grade ITBS as a premeasure.

Wiggins (1989) all argue that it is natural for teachers to work hard to prepare their students to do well on examinations that matter. Rather than forbid "teaching to the test," which is impossible, it is preferable to create measures that will result in good instruction even when teachers do what is natural. The reshaping of instruction toward desirable processes and outcomes is expected to occur both indirectly, as teachers individually imitate assessment tasks in a variety of ways, and directly, because criteria for judging performances will be shared explicitly.

These anticipated benefits of performance assessments are conceptually grounded and supported by evidence documenting the negative effects of traditional, standardized testing. Under conditions of high-stakes accountability pressure, it has been demonstrated that teachers align instruction with the content of basic skills tests, often ignoring science and social studies and even untested objectives in reading and mathematics. Furthermore, instruction on tested skills comes to resemble closely the format of multiple-choice tests, with students learning to recognize right answers rather than generating their own problem solutions (Madaus, West, Harmon, Lomax, & Viator, 1992; Shepard, 1991; Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990). Such measurement-driven instruction has been harmful to learning as evidenced by the decline in higher order thinking skills on the National Assessment of Educational Progress during the 1980s and by the failure of accountability test score results to generalize when students are retested using less familiar formats (Flexer, 1991; Hiebert, 1991; Koretz, Linn, Dunbar, & Shepard, 1991).

Thus it is the obverse case that has been "proven." Teaching to standardized tests harms both teaching and learning. Advocates of performance assessments assume, therefore, that parallel mechanisms will work to produce positive effects once limited tests are replaced by more desirable measures. However, to date little research has been done to evaluate the actual effects of performance assessments on instructional practices or on student learning. Although some extreme views hold that authentic performance measures are valid by definition and will automatically produce salutary effects, we would argue in contrast that the effects of performance assessments should be evaluated empirically following a program of inquiry closely parallel to the studies undertaken to examine the effects of standardized tests. We concur with Linn, Baker, and Dunbar (1991) that validity criteria for alternative assessments should address intended and

unintended effects as well as more substantive features such as cognitive complexity, content quality and comprehensiveness, generalizability of knowledge from assessed to unassessed tasks, and the like. Although we are committed to performance assessments on conceptual grounds, their demonstrated effects on teaching and learning remain an open question.

The purpose of the present study was to examine the effects of performance assessments on student learning. If standardized tests are removed and teachers begin to use performance assessments as part of regular instruction, will student performance on independent measures of achievement be improved? Note that some arguments favoring the use of performance assessments to leverage educational reform presume that the high-stakes accountability pressures would still be needed to drive instructional change. Other advocates focus more on the informational and feedback effects of classroom-embedded assessments. In this study, we adopted the second perspective. We were interested in the effects of using assessments as part of instruction but without the incentives and context created by an externally mandated system.

A year-long project was undertaken to help teachers in 13 third-grade classrooms begin to use performance assessments as a part of regular instruction in reading and mathematics. Other parts of the research project focused on changes in teachers' beliefs and practices about summaries and expository text in reading (Borko, Davinroy, Flory, Hiebert, 1994; Davinroy & Hiebert, 1993); on changes in teachers' beliefs about assessment and instruction in mathematics (Flexer, Borko, Cumbo, Marion, & Mayfield, 1994); on parent attitudes toward performance assessments (Shepard & Bliem, 1993); and on student understandings of how teachers "know what they know" (Davinroy, Bliem, & Mayfield, 1994). Here research questions are focused on student achievement in reading and mathematics. Did students learn more or develop qualitatively different understandings because performance assessments were introduced into classrooms? Achievement results were compared both to the performance of third-grade students in the same schools the year before and to third-grade performance in matched control schools.

Study Methods

Setting

The study was conducted in a working-class and lower-to-middle-class school district on the outskirts of Denver, Colorado. The district was selected in part because of the willingness of central office administrators to participate and in part because of its ethnically diverse student population. In the 1980s the district was known for its extensive mastery learning and criterion-referenced testing system, but more recently curriculum guidelines in language arts and mathematics were revised to reflect more constructivist conceptions of these disciplines, consistent with national standards (Anderson, Hiebert, Scott, & Wilkinson, 1985; National Council of Teachers of Mathematics, 1989).

We wanted teachers to be free to implement performance assessments and to make concomitant changes in instruction without worrying about how their students would do on the standardized test normally administered every April. Therefore, a requirement of participation was that the district be willing to apply to the state for a 2-year waiver from standardized testing in the three schools selected to participate. As part of its procedures to grant the waiver the state required, in turn, that approvals be obtained from the school board, the district accountability committee, the teachers' union, and parent accountability committees in each of the participating schools.

Sample and Research Design

Third grade was selected as the target grade level because district CTBS testing occurs at Grades 3 and 6 (but not all sixth grades are in elementary schools). Because of the amount of time and effort that would be required of teachers, volunteer schools were sought. Third-grade teachers had to make a commitment as a team with the support of their principal and parent accountability committee. Ten schools sent representatives to a workshop where the study's purpose and methods were explained. We accepted the three schools that completed the formal application to participate. In the 1992-93 study year there were 13 third-grade classrooms in the three schools combined involving approximately 335 third graders.

Three control schools were identified to be used for comparison when analyzing teachers' beliefs and parents' opinions as well as students' achievement. The control and participating schools were matched on free and reduced lunch

percentages, percentage of minority children, and other knowledge of neighborhood similarities such as type of housing. Data in Table 1 show the socioeconomic differences among the three participating schools as well as their matches to control schools. Note that the implementation year of the project was 1992-93. Therefore, site selection and baseline testing occurred in the spring of 1992; data available for school matching had been gathered spring 1991.

CTBS achievement test data for participating and control schools are shown separately in Table 2. As part of the matching process, we found that it was impossible to match schools on both socioeconomic factors and 1991 CTBS scores because they diverged too much. This was unusual. In our experience in other studies, test scores and socioeconomic indicators usually correspond closely enough that it is possible to select schools that are the same on both. Because we could not know whether sharp differences in achievement scores meant more able populations, more able teaching, or more test-score inflation in the candidate control schools, we elected to match only on socioeconomic data. However, subsequent to the selection process, we administered our own baseline achievement measures in reading and mathematics which confirmed the superior performance of third graders in control schools in the year before the study began.

The research design called for two separate comparisons. Outcome measures in reading and mathematics selected for administration in May 1993 were also administered as baseline measures in May 1992. In addition, premeasures appropriate to entering third graders were administered in September 1992 and used as covariates to evaluate 1993 outcomes.

Assessment Project "Intervention"

The intention of the project was not to introduce an already-developed curriculum and assessment package. Rather, we proposed to work with teachers to help them develop (or select) performance assessments congruent with their own instructional goals. Faculty researchers included Roberta Flexer, an expert in mathematics, Elfrieda Hiebert, an expert in reading, Hilda Borko, whose specialty is teacher change, and Lorrie Shepard, an assessment expert. We met with teachers for planning meetings in spring 1992 and September 1992. Then we met for weekly afterschool workshops for the entire 1992-93 school year, alternating between reading and mathematics so that subject matter specialists could rotate among schools.

Table 1

Socioeconomic Characteristics of Participating and Control Schools

	Participating schools			Control schools		
	1	2	3	1	2	3
Free and reduced lunch	61%	9%	6%	55%	13%	3%
Percent minority	37%	16%	14%	45%	19%	10%
Student turnover	27%	7%	11%	30%	11%	10%

Table 2

Grade 3 Mean CTBS Scores in Reading and Mathematics for Participating and Control Schools

	Participating schools			Control schools		
	1	2	3	1	2	3
5-Year Average (1987-91)						
Total reading	47.8	48.8	52.7	48.9	50.4	54.7
Total mathematics	52.5	47.5	51.3	49.3	60.9	58.1
1991						
Total reading	47.0	43.0	47.0	47.0	54.0	56.0
Total mathematics	54.0	52.0	50.0	50.0	67.0	63.0
1992						
Total reading	44.6	51.9	55.5	43.1	54.4	57.2
Total mathematics	53.9	53.8	62.8	47.5	66.5	68.1
1993						
Total reading	N/A[a]	N/A[a]	N/A[a]	49.6	47.0	57.2
Total mathematics	N/A[a]	N/A[a]	N/A[a]	57.1	57.1	65.3

[a] Participating schools were exempt from CTBS testing in 1993 as a condition of the study.

Because the district had newly developed curriculum frameworks consistent with emerging national standards in reading and mathematics, and because teachers had volunteered to participate in the project, we assumed that their

views about instruction would be similar to those reflected in the district curriculum and therefore similar to our own. What we learned later was that not all teachers were true volunteers; some had been "volunteered" by their principals or had acquiesced to pressure from the rest of the third-grade team. More importantly, for understanding the substantive character of the project, even some teachers who were willing and energetic project participants were happy with the use of basal readers and chapter tests in the math text and were not necessarily familiar with curricular shifts implied by the new district framework in mathematics.

Although dissonance between researchers' and teachers' views about subject matter instruction was sometimes acknowledged and joked about in workshops, for the most part researchers avoided confrontations about differences in beliefs and did not propose radical changes in instruction. Faculty experts worked to suggest possible reading and mathematics activities that addressed teachers' goals but that departed from a strictly skills-based approach. For example, we refused to consider having timed tests on math facts as part of project portfolios, but in other ways we conformed to teacher-identified goals.

At the start of the year, teachers said that they wanted to "start small" but then selected as goals meaning-making and fluency in reading, and understanding of place value, addition and subtraction, and multiplication, as the foci for the project. In the fall, for reading, teachers learned to use running-records to assess fluency for below-grade-level readers. Written summaries were used to assess comprehension; but for some teachers summaries became an end in themselves (Borko et al., 1994). Project activities included the development of rubrics to score written summaries. In the spring, ideas about meaning-making and written summaries were extended to expository texts.

In mathematics, teachers made extensive requests, throughout the year, for materials and ideas for teaching the topics of the third-grade curriculum, for example, place value, addition, geometry, and probability. Materials that addressed these topics from a problem-oriented and hands-on approach were distributed to all three schools to use in both instruction and assessment. Teachers were offered nonroutine problems from which to select a number to try with their classes. Some problems required students to explain their solutions; others required students to analyze and explain an incorrect step or computation in a buggy problem. Materials were also distributed for making and using base 10

blocks for modeling numbers and operations. Some teachers had not previously worked with place-value mats or manipulatives and introduced them for the first time. Discussions at weekly meetings included dialogue about using materials for both instruction and assessment, making observations and how to keep track of them, analyzing student work, and developing rubrics for scoring problem solving and explanations.

Outcome Measures and Covariates

For obvious reasons, we did not wish to use a multiple-choice standardized test to measure the project's effects. At the same time, a compendium of performance tasks used throughout the project would also not be a fair outcome measure. We are grateful to the Maryland State Department of Education for allowing us to use portions of their 1991 performance assessments in reading and mathematics as outcome measures for the study. Although the Maryland assessments are still relatively test-like compared to week-long projects that students might do, they are markedly different from traditional tests. The tasks provide sufficient structure and support so that students in the baseline year and in control schools could understand what they were being asked to do, but they are sufficiently open-ended that students had to produce answers to show what they knew. In literacy, students read extended stories and informational texts in a separate reading book and then wrote responses about what they read, completed tables, drew story webs, and so forth. In mathematics, tasks involved a series of problems all related to the same information source or application. Students had to solve problems that involved identifying patterns, estimating as well as computing, using calculators, extending tables, and explaining how they got their answers. Because we were limited to only four 1-hour sessions to administer our outcome measures, we used only a sample of tasks from the Maryland assessments.

We wanted to be sure to assess a range of skills in mathematics. Therefore, we used three tasks from the Maryland assessment in one 1-hour session, but also used a portion of an alternative measure in mathematics developed for another study (Koretz et al., 1991). This test consisted of 15 short-answer and multiple-choice items that assess problem solving in, and conceptual understanding of, functions and relations, patterns, whole-number operations, probability, and data and graphs. Problem-types included application and nonroutine problems.

Covariate measures were needed for entering third graders to assess their initial abilities in reading and mathematics. In reading, portions of a Silver, Burdett and Ginn 2/3 Reading Process Test and 2/3 Skills Progress Test were used with permission from the publisher. After reading a 13-page story-and-pictures book, students responded to questions by checking answers (more than one answer could be correct) and also writing responses. Students also read two page-long passages and responded to comprehension questions by circling the correct answer. In mathematics, open-ended problems were developed to measure students' ability to discern patterns and number relations. This subtest was combined with three subtests from the second-grade level of the Iowa Tests of Basic Skills covering math concepts, estimation, and data interpretation. We are grateful to Riverside Publishing Company for allowing us to reproduce portions of their test for use in the study. The reading and math covariates were each administered in 1-hour sessions on separate days.

Scoring and Reliability

All of the measures used in the study required scoring of open-ended student responses. In particular, the Maryland assessment tasks required scorers to make subjective judgments about the quality of student answers. Therefore, these instruments received the greatest scrutiny in our reliability studies. Scorers worked from the scoring guides provided by the Maryland School Performance Assessment Program with slight modifications made by the respective subject matter experts. Day-long training sessions were held in summer 1992, and again in 1993, to ensure that scorers were familiar with the scoring rules and able to apply them to the full range of students' responses.

Interrater reliability was assessed both within year (are all of the scorers rating consistently?) and between years (were the scoring rules implemented consistently in 1992 and 1993?). For the within-year studies, three student booklets in reading and three in mathematics were chosen at random from each classroom. This resulted in more than a 10% sample with 55 to 60 out of 500 booklets being rescored. Booklets were scored independently by the scorer-trainer. Three other raters were then compared one at a time and then in aggregate to this standard rater. Pearson correlations between total scores assigned by other raters and by the standard rater were quite high in both years for both reading and mathematics; values ranged from .96 to .99. The Maryland reading measure was composed of 61 scored "items" or task subparts; the Maryland mathematics

measure had 31 scorable entities. The high correlations between raters simply mean that with sufficient numbers of task subscores, raters can rank students quite accurately.

A truer picture of the effect of rater agreement on total scores is provided by the data in Tables 3 and 4. On individual items requiring a subjective judgment, raters might differ by only one point in how they scored the item. However, these discrepancies could accumulate across items. The data in Tables 3 and 4 show how often raters agreed completely with the standard rater on total score and how often they differed by four or more points in reading or two or more points in mathematics. Within years, raters agreed on total score within one or two points for 97% or 98% of cases in reading and for 90% to 91% of cases in mathematics. These agreement rates are respectable for subjectively scored instruments but nonetheless introduce noise into the evaluation of effects.

To check for consistency of scoring across years, test booklets from 1992 were "seeded" into 1993 classroom sets without scorers being aware of which booklets were being rescored. A total of 57 booklets were rescored in both mathematics and reading. As seen in Tables 3 and 4, the between-year agreements were not so high as the within-year agreements. In mathematics, 79% of total scores were within 2 points of the score assigned to the same booklet the year before. In reading, 72% were within 4 points (which is comparable in standard deviation units to a 2-point difference on the mathematics assessment). The between-years analysis also revealed some systematic biases with raters tending to become more stringent in 1993 than raters had been in 1992. In reading there was an average mean score shift downward for the 57 1992 booklets rescored in 1993 of 2.47 points. In math the greater stringency created a downward shift of .25 points. Because the reading score shift was both statistically and practically significant, 1993 reading scores were adjusted to correct for the systematic bias. Average biases varied for individual raters from 1.13 to 3.63, all in the direction of greater stringency; these specific corrections were applied to the sets of booklets scored by each rater.

Internal consistency coefficients provide another indicator of the psychometric adequacy of research instruments. Coefficients calculated on the entire sample are shown in Table 5 for the covariates and for both the 1992 and

Table 3

Percentage of Scorer Agreement on Maryland Reading Assessment Total Score

	Within year 1992	Within year 1993	Between year 1992/1993
Complete agreement	33.3%	28.6%	12.2%
Agreement within ± 2 points (± .17 *SD*)	97.8%	96.5%	45.6%
Agreement within ± 4 points (± .34 *SD*)	100.0%	100.0%	71.9%
Range of differences	-2% to +4%	-3% to +2%	-5% to +9%

Note. Agreement is based on the comparison between each rater's judgment of total student score and the independent rater's judgment of student scores. ± 4 points was used in the reading analyses because the total number of possible points was 61 compared to 31 in the mathematics assessment. In standard deviation (*SD*) units, differences of 2 and 4 points in reading are roughly comparable to 1 and 2 points in mathematics.

Table 4

Percentage of Scorer Agreement on Maryland Mathematics Assessment Total Score

	Within year 1992	Within year 1993	Between year 1992/1993
Complete agreement	30.4%	29.0%	14.3%
Agreement within ± 1 point (± .15 *SD*)	75.0%	64.5%	54.0%
Agreement within ± 2 points (± .31 *SD*)	91.0%	90.3%	79.4%
Range of differences	-4% to +3%	-4% to +3%	-3% to +4%

Note. Agreement is based on the comparison between each rater's judgment of total student score and the independent rater's judgment of student scores.

1993 administrations of the outcome measures. Although low coefficients could mean either poor reliability or task-item heterogeneity, high values provide assurance that summary scores are reliable and reasonably consistent measures of student performance.

Table 5

Internal Consistency Coefficients (Cronbach's Alpha) for Outcome and Covariate Measures

	1992		1993	
	n	alpha	*n*	alpha
Maryland reading	458	.90	458	.90
Covariate reading		n/a	458	.74
Maryland mathematics	487	.84	523	.83
Alternative mathematics	487	.78	524	.80
Covariate mathematics		n/a	454	.85

Results

Outcome measures for 1993 were analyzed in two ways, first in comparison to 1992 baseline administrations of the same measures, and then in relation to control group outcomes using analysis of covariance. To make it easier to follow the logic of the two analyses, results are reported separately in Tables 6 and 7. The 1993 data are repeated in both tables, although subjects without pretest data were deleted from the analysis of covariance (Table 7). Then the analyses are presented graphically in Figures 1 and 2 for reading and mathematics respectively.

Overall, the predominant finding is one of "no-difference" or no gains in student learning following from the year-long effort to introduce performance assessments. Although we argue subsequently that we see qualitative changes in student performance and that the small gain in mathematics is real, honest discussion of project effects must acknowledge that any benefits are small and ephemeral. For example, improvements show up in some project-teachers' classrooms but not in all, and the largest gain from 1992 to 1993 for the participating schools combined, which occurred on the Maryland mathematics assessment, produced an effect size of only .13.

Table 6

1992 vs. 1993 Comparisons in Reading and Mathematics for Participating and Control Schools

	1992 Mean (*n*)	1993 Mean (*n*)	92-93 Mean difference	1992 Pooled w/in school *SD*	ES[a] of difference
Maryland reading total					
Participating	27.7 (290)	26.1 (305)	-1.6		-.14
				11.7	
Control	28.9 (210)	26.5 (228)	-2.4		-.21
Maryland math total					
Participating	12.2 (288)	13.0 (305)	0.8		.13
				5.94	
Control	15.3 (210)	13.6 (231)	-1.7		-.29
Alternative math total					
Participating	12.7 (288)	12.9 (305)	0.2		.06
				3.5	
Control	13.3 (208)	13.5 (229)	0.2		.06

[a] Effect size calculations are based on pooled within-school 1992 standard deviations using both participating and control group schools.

In reading there were no significant differences between 1992 and 1993 results or between participating and control schools. Both groups of schools appeared to lose ground slightly (.9 and 1.9 points respectively).

In mathematics the alternative test also showed no effects. However, the Maryland assessment in mathematics, which requires students to do more extended problems and explain their answers, did show a positive gain in the participating schools. We interpret this change, albeit small, as a "real" gain based on the following arguments. First, CTBS results for 1993 showed declines districtwide and in two of the control schools. Against a backdrop of declining achievement, slight gains in the participating schools are more impressive. Although the populations of the participating and control schools are quite similar as evidenced by socioeconomic variables and pretest measures, third graders in the control schools have traditionally outperformed third graders in the participating schools. This was apparent in five years of CTBS data and

Table 7

1993 Outcome Comparisons Between Participating and Control Schools With and Without Covariance Adjustments

	1993 Mean	Sept. 1992 pretest[a]	May 1993 adjusted mean[b]
Maryland reading total			
Participating	26.8	11.7	26.2
Control	27.0	10.8	27.9
Difference	-0.2	0.9	-1.7
Maryland math total			
Participating	13.0	19.8	13.1
Control	13.9	20.4	13.8
Difference	-0.9	-0.6	-0.7
Alternative math total			
Participating	13.0	19.8	13.0
Control	13.8	20.4	13.7
Difference	-0.8	-0.6	-0.7

[a] There was one mathematics pretest and one reading pretest (different from the Maryland assessment or the alternative assessment); the pretest scores are repeated with each measure for ease of reference.

[b] The "1993 adjusted means" are the 1993 mean scores statistically "adjusted" for the September 1992 pretest scores.

on the 1992 baseline measure in mathematics. Therefore, one way of interpreting the between-year and covariance analyses together is to say that the assessment project helped participating students "catch up" to the control students in math achievement. From all indications, this would not have occurred without the project.

We also noted qualitative changes in students' answers to math problems which suggest that at least in some project classrooms whole groups of students were having opportunities to develop their mathematical understandings that had not occurred previously. Figure 3 and Tables 8 and 9 were constructed to provide a qualitative summary of student responses to a sample problem and to illustrate what slight improvements in student scores mean substantively. The two classrooms that showed the greatest gains from 1992 to 1993 in the low socioeconomic participating and control schools were selected for comparison (Table 8). Both teachers' classrooms showed an effect-size gain of .27 from 1992

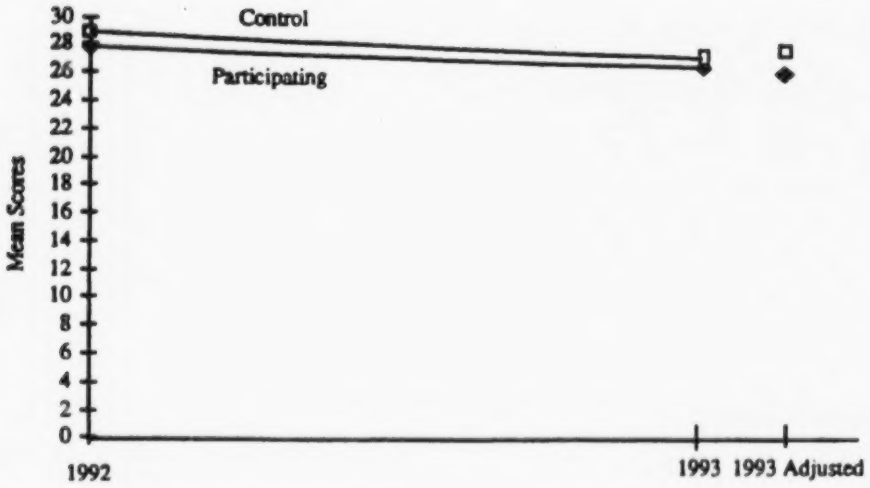

Figure 1. Maryland Reading Assessment mean scores for participating and control schools.

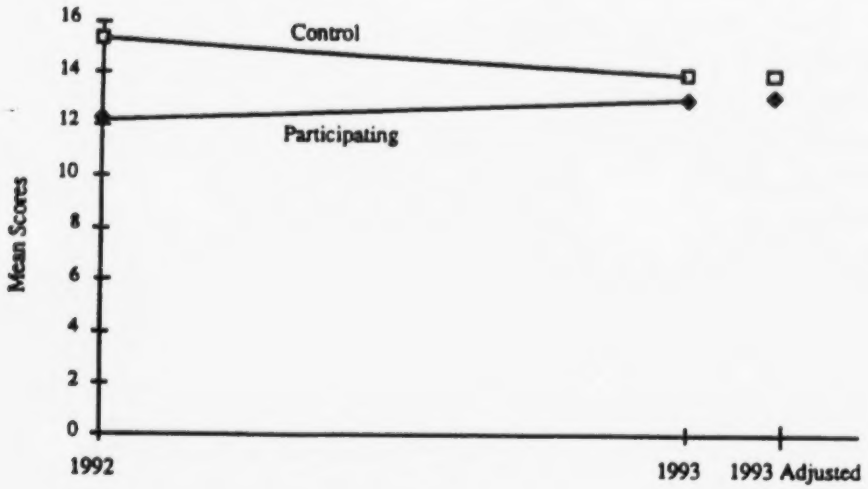

Figure 2. Maryland Mathematics Assessment mean scores for participating and control schools.

BEST COPY AVAILABLE

to 1993 on the Maryland Mathematics Assessment. However, for this particular problem we think we see a greater gain for the participating classroom—one that suggests that a whole classroom of typically poorly performing students had developed knowledge of patterns and mathematical tables that this teacher's students had not understood the previous year. At the top of the scale there are no more right answers in 1993 than in 1992. However, in 1993 84% of the children in the participating classroom could complete the table (Categories I-V), whereas in 1992 only 34% of the same teacher's students could complete this part of the problem. The percentage of students in the participating classroom who could write explanations describing a mathematical pattern or telling how they used the table (Categories I, III, or IV) also increased substantially, from 13% to 55%. Even students who took the *wrong* answer from their table could describe the pattern:

- I counted by fours which is 60 the(n) I went in the ones which is 15.
- I counted by 4 and ones and came to 60.
- First I went up to 15 pitchers. Then I made 60 cups.
- (60) first I cont'd by one's then I contid by fors
- first I saw that the where counting by 4s So I counted by fours. until there was no rome and got the answer 57.
- (15) I counted by 4s and I lookt at the top one

In the control low SES classroom the percentage of students writing explanations actually declined from 39% to 23%. Obviously for the two teachers to have the same positive gain in total score, there must be other problems where the control class gained relatively more. However, the qualitative analysis suggests that the gains in the control classroom tended to be more randomly distributed, with individual children earning a few more points here and there, but with no systematic shifts like the one just described. We are more inclined to attribute changes of this type to changes in instruction.

In Table 9, 1992 versus 1993 comparison data are shown for the same problem but for the two "best" classes in the highest socioeconomic pair of schools. Note that selecting the best class in the control school meant selecting the class with the smallest decline (ES = -.20) on the Maryland Mathematics

I. Extends Table, Answers correctly, Explains either pattern or point in chart.

STEP 4

Now you want to know how many pitchers you will need for 46 cups of lemonade. You can see from the table below that a one-quart pitcher will hold 4 cups, and 2 one-quart pitchers will hold 8 cups. Continue the pattern in both rows of the table until you find the number of pitchers needed to hold 46 cups of lemonade.

Pitchers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cups	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60

How many one-quart pitchers will you need for 46 cups of lemonade? Write your answer on the line below.

12

Explain how you got your answer. Write on the lines below.

I looked at the patteren and saw that there was not a 46. so I took 48 so there is also some for my friend and I.

Explain how you got your answer. Write on the lines below.

From pichers #11 to 12 it went 44 48 cups so I just put 11 to

Figure 3. Sample student responses on Maryland Mathematics Assessment problem set two (Lemonade Step 4) illustrating key qualitative categories.

IV. Extends table, Wrong Answer (60, 15, 11, other), Explanation describes pattern.

Pitchers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cups	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60

How many one-quart pitchers will you need for 46 cups of lemonade? Write your answer on the line below.

15

Explain how you got your answer. Write on the lines below.

I counted by fours which is 60 the I went in the ones which is 15.

Explain how you got your answer. Write on the lines below.

On the cups as you go along you count four more each time

Explain how you got your answer. Write on the lines below.

first I saw that the where counting by 4's So I counted by fours. until there was no remaind got the answer 15.

Figure 3 (continued).

Table 8

Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) From the Classrooms With the Greatest Gains in Low Socioeconomic Participating and Control Schools

		Participating		Control	
		1992	1993	1992	1993
I.	Extends table, Answers correctly, Explains (explains either pattern or point in chart).	13%	13%	31%	19%
II.	Extends table, Answers correctly, Inadequate explanation.	4%	0	8%	12%
III.	No answer but stops table at right place, Explanation describes pattern.	0	0	0	0
IV.	Extends table, Wrong answer (60, 15, 11, other), Explanation describes pattern.	0	42%	8%	4%
V.	Extends table, Wrong answer (60, 15, 11, other), Inadequate explanation.	17%	29%	8%	35%
VI.	Cannot extend table.	63%	8%	46%	31%
VII.	Blank.	4%	8%	0	0

Table 9

Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) From the Classrooms With the Greatest Gains in High Socioeconomic Participating and Control Schools

		Participating		Control	
		1992	1993	1992	1993
I.	Extends table, Answers correctly, Explains (explains either pattern or point in chart).	19%	43%	56%	43%
II.	Extends table, Answers correctly, Inadequate explanation.	8%	0	0	4%
III.	No answer but stops table at right place, Explanation describes pattern.	0	5%	0	0
IV.	Extends table, Wrong answer (60, 15, 11, other), Explanation describes pattern.	12%	29%	39%	9%
V.	Extends table, Wrong answer (60, 15, 11, other), Inadequate explanation.	31%	9%	0	30%
VI.	Cannot extend table.	31%	9%	6%	13%
VII.	Blank.	0	5%	0	0

Assessment, given that all classrooms in this school started higher in 1992 than any other classrooms but declined slightly in 1993. In contrast, the best classroom in the matched participating school showed a substantial improvement (ES = .53) and caught up to where the best control classrooms had been the year before.

Although the level of student performance is much higher in Table 9 than in Table 8, corresponding to the difference in socioeconomic level of the schools, the participating classroom in Table 9 still shows specific improvements in student performance that can be associated with the project intervention. Of course, there are more right answers (Category I), 43% versus 19% in 1992. More importantly, however, in 1993 77% of the children in the participating classroom wrote mathematically adequate explanations (Category I, III, or IV) about how they solved the problem. This proportion is in contrast to 31% who wrote explanations in the same teacher's classroom the year before. In the control classroom, 95% wrote adequate explanations in 1992 but only 52% could do so in 1993. As explained previously, we are more inclined to attribute these declines to population changes rather than to a decline in the quality of teaching, especially because all classrooms in the control school were affected. Table 9 also illustrates the increased ability of students in participating schools to extend a mathematical pattern or complete a function table. In 1992 only 70% of the participating children could extend the table (Categories I-V), but this percentage increased to 86% in 1993 making the participating classroom more comparable to the high levels achieved in the control classroom both years (95% and 86%, respectively).

Samples of student responses to a different problem or portion of the lemonade task are presented in Figure 4. Again we have chosen to illustrate the qualitative categories where students wrote explanations; these answers received either whole or partial credit in the quantitative scoring. This problem was much more difficult for children across schools and did not show much of an improvement for the low socioeconomic best classroom. There were no more right answers than in 1992, but 27% of students wrote mathematically adequate descriptions of the pattern (Category V) compared to 0% in 1993. However, a similar improvement occurred in the low SES matched classroom.

I. Right answers, Explanation describes pattern (includes minimal explanation 6 + 6 = 12).

You and your friend are in charge of preparing lemonade for 2 classes. You must decide how much lemonade to make for 46 students. Each student should get a cupful of lemonade.

STEP 1 Read this table from a lemonade mix container.

Scoops	Cups Made
1	2
3	6
5	10

7 14
9 18
11 22
13 26
15 30
17 34
19 38
21 42
23 46

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

46
÷ 2
23

If you see ☆ in the table you can make half as many with the shapes so the ansar is 23

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23!!!!

Figure 4. Sample student responses on Maryland Mathematics Assessment problem set one (Lemonade Steps 1-2) illustrating key qualitative categories.

I. Right answers, Explanation describes pattern (includes minimal explanation 6 + 6 = 12).
Additional Examples

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

yw old the same neber
so exopel (6+6=12) (6X2=12)

If you put 1 scoop it will make 2. Then if you have 3 scoops it will make 6. So every scoop you do you will have to dubble that number

With 6 scoops of mix you should be able to make 12 cups of lemonade I figred this out because 1+1=2, 3+3=6, 5+5=10 so 6+6=12 so that means you have 12 full cups of lemonade.

23
23
+
46

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

Figure 4 (continued).

V. Attempts to extend table but focuses on Left or Right column, not Left:Right pattern OR sees 1:2 pattern but can't apply to get answers.

STEP 1 Read this table from a lemonade mix container.

Scoops	Cups Made
1	2
3	6
5	10
7	14
9	18
11	22
13	26
15	30
17	34
19	38
21	42
23	46

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

because on scoops it's go 1,3,5 I saw that their doing all odd so I put odd why cups was all even and 4 in the mitel. What I mean is 2 + 4 = 6 and 6 + 4 = 10 and so on.

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

Figure 4 (continued).

Category V responses show some of the richness of the students' answers and also help us to understand why many students found this problem more difficult. In every classroom there were some students who could count by fours when they got to step 4 but had trouble with steps 1-2 because they extended the table downward without looking at the left-right correspondence. They were able to explain what they were thinking mathematically in a way, in fact, that revealed their misconception:

- Yes I do see a pattern, on the side with the spoon it counts by 2's were there's a cup it counts by fours.
- because on scoops it's go 1, 3, 5, I saw that their doing all odd so I put odd why cups was all even and 4 in the mitel. What I mean is 2 + 4 = 6 and 6 + 4 = 10 and so on.

The high SES classroom did show a substantial gain on this problem; the data for the comparison participating and control classroom are shown in Table 10. From 1992 to 1993 the percentage of students who wrote mathematical explanations (and extended the table) increased from 27% to 57% (Categories I, III, V). The corresponding change in the control classroom was from 45% to 35%.

The qualitative analyses of student answers on the Maryland Mathematics Assessment afforded us an appreciation of what an effect size of .13 means in terms of substantive increases in student learning. Significant shifts in specific categories occurred in participating classrooms but not in control classrooms and are associated with the kinds of mathematical activities introduced as part of the project. In many cases this meant that students in the middle and bottom of the class were able to do things that their counterparts had not been able to do the previous year. These changes were confirmed quantitatively by significant shifts out of the lowest two quintiles for two of the three participating schools; (cut scores for quintiles were defined in 1992 on the total sample and held constant for 1993).

Table 10

Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set One (Lemonade Steps 1–2) From the Classrooms With the Greatest Gains in High Socioeconomic Participating and Control Schools

		Participating		Control	
		1992	1993	1992	1993
I.	Right answers, Explanation describes pattern (includes minimal explanation, 6 + 6 = 12).	19%	24%	39%	9%
II.	Right answers, No explanation (but may show 23 + 23 = 46).	0	0	6%	9%
III.	Gets 12 cups with adequate explanation but cannot extend to 46 cups.	8%	9%	0	0
IV.	Gets 12 cups, Inadequate explanation, (wrong or no extension).	4%	0	0	4%
V.	Attempts to extend table but focuses on L or R column, not L/R pattern, OR 1:2 correspondence without answers, Explains thinking.	0	24%	6%	26%
VI.	Wrong answers, Explanation not based on chart or only restates answer.	58%	9%	33%	48%
VII.	Wrong answers, No explanation.	4%	29%	0	0
VIII.	Blank.	8%	5%	17%	4%

Conclusions

A fairly elaborate research design was implemented to evaluate the effect of a year-long performance assessment project on student learning. Maryland third-grade assessments in reading and mathematics and another alternative mathematics test served as independent measures of student achievement, separate from the classroom assessments developed as part of the project. End-of-year results in 1993 were compared to baseline administrations of the same measures in 1992 and to control school performance using analysis of covariance.

Results in reading showed no change or improvement attributable to the project. Third-graders in the participating schools did about the same on the Maryland Reading Assessment as third-graders had done the year before, and there were no significant differences between participating and control schools. In mathematics there were also no gains on the alternative assessment measure.

However, small quantitative changes and even greater qualitative changes did occur on the Maryland Mathematics Assessment.

It is possible to offer both pessimistic and optimistic interpretations of the study results. Most significantly, from a negative perspective, it is clear that introducing performance measures did not produce immediate and automatic improvements in student learning. This finding should be sobering for advocates who look to changes in assessment as the primary lever for educational reform.

Of course, there were mitigating factors that help to explain and contextualize the lack of dramatic effects. First, we did not "teach to" the project outcome measures. For example, the classroom use of written summaries to assess meaning-making should have given students more experience with certain open-ended responses on the Maryland Reading Assessment. However, we did not introduce any other item formats from the outcome measure such as comparative charts or story webs. We should also note that the level of text difficulty in the Maryland assessment was quite high. In retrospect, we might have included additional, easier texts to be more sensitive to gains by below-grade-level readers.

Similarly, in mathematics we worked on explanations and used function tables as one of several problem-solving strategies (along with "guess and check," draw a picture, and "use cubes" [make a model]) but did not use formats that conformed specifically to the Maryland assessment. It is reasonable to assume that teachers might have behaved differently and imitated the outcome measures more closely if our 1992 baseline administration and anticipated 1993 measure had been imposed by an external agency for accountability purposes. Such practices could very likely have heightened the improvement of outcome scores, but then the question would arise as to whether the increased scores validly reflected improvement in students' understanding.

When we showed project teachers the outcome findings (in fall 1993), they were disappointed but offered an explanation regarding the "intervention" that jibes with our own sense of the project's evolution. Despite the level of workshop effort throughout 1992-93, by Christmas project "assignments" still had not been assimilated into regular instruction. Although we have evidence of changes beginning to be made in the spring term (Flexer et al., 1994), many teachers said that they did not "really" change until the next year (1993-94) (beyond the reach

of the outcome measures). Several teachers argued that they did not fully understand and adopt project ideas and assessment strategies until they began planning and thinking about what and how to teach this year. This view is consistent with the literature on teacher change. Fundamental and conceptual change occurs slowly. Furthermore, changes in student understandings must necessarily come last, after changes in teacher thinking and changes in instruction.

We also note that the apparent gain in mathematics compared to zero gain in reading might have occurred because teachers had "further to go" in mathematics than in reading. If we take district curriculum frameworks as the standard, which are consistent with emerging professional standards in the respective disciplines, most teachers in the participating schools had already implemented some instructional strategies focused on meaning-making. In mathematics, the district frameworks were newer, and teachers were less familiar with them. Two teachers had tried out the Marilyn Burns (1991) multiplication unit the year before; but several more teachers decided to try it during the project year. Several were using manipulatives for the first time; several adopted materials to teach problem-solving strategies for the first time; and one group of teachers worked to develop new units in geometry and probability. Even when teachers did not understand them well or use materials optimally, these brand-new activities represented substantial shifts in the delivered curriculum.

In contrast to these apologies and caveats about why change did not occur, the cause for optimism comes from the small but real gains in mathematics. Because of the project, most of the teachers in the participating schools spent class time on written explanations (especially what makes a good explanation) and on mathematical patterns and tables, which they had never done before. As a consequence, there were specific things that a large proportion of third graders in these classrooms could do on the outcome assessments, where before only the most able third graders had been able to intuit how to do them.

Our concluding advice is that reformers take seriously the current rhetoric about "delivery standards" and the need for sustained professional development to implement a thinking curriculum. Performance assessments—even with the diligent effort of most project teachers and the commitment of four university researchers—did not automatically improve student learning. The changes that did occur, however, confirm our beliefs that many more students can develop

conceptual understandings presently exhibited by only the most able students—if only they are exposed to relevant problems and given the opportunity to learn. Performance assessments that embody important instructional goals are one way to invite instructional change, and assessments have the added advantage of providing valuable feedback about student learning. However, we would not claim that performance assessments are necessarily the most effective means to redirect instruction. When teachers' beliefs and classroom practices diverge from new conceptions of instruction, it may be more effective to provide staff development to address those beliefs and practices directly. Performance assessments are a key element in instructional reform, but they are not by themselves an easy cure-all.

References

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading.* Champaign, IL: The Center for the Study of Reading, National Institute of Education, National Academy of Education.

Borko, H., Davinroy, K. H., Flory, M. D., & Hiebert, G. H. (1994). Teachers' knowledge and beliefs about summary as a component of reading. In R. Garner & P. A. Alexander (Eds.), *Beliefs about texts and instruction with text* (pp. 155-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Burns, M. (1991). *Math by all means: Multiplication grade 3.* Sausalito, CA: The Math Solution Publications.

Davinroy, K. H., Bliem, C. L., & Mayfield, V. (1994, April). *"How does my teacher know what I know?": Third-graders' perceptions of math, reading, and assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Davinroy, K. H., & Hiebert, E. H. (1993, December). *An examination of teachers' thinking about assessment of expository text.* Paper presented at the annual meeting of the National Research Conference, Charleston, SC.

Flexer, R. J. (1991, April). *Comparisons of student mathematics performance on standardized and alternative measures in high-stakes contexts.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Flexer, R. J., Borko, H., Cumbo, K., Marion, S., & Mayfield, V. (April, 1994). *How "messing about" with assessment affects instruction.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Hiebert, E. H. (1991, April). *Comparisons of student reading performance on standardized and alternative measures in high-stakes contexts.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12: Executive summary.* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 72*, 232-238.

Shepard, L. A., & Bliem, C. L. (1993, April). *Parent opinions about standardized tests, teachers' information and performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1990). *The role of testing in elementary schools.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.